Origins of power-law degree distribution in the heterogeneity of human activity in social networks

Lev Muchnik[1*], Sen Pei[2,3*], Lucas C. Parra[4*], Saulo D. S. Reis[2,5],
José S. Andrade, Jr.[5], Shlomo Havlin[6], and Hernán A. Makse[2,5]

[1]*School of Business Administration,*
The Hebrew University of Jerusalem, 91905 Israel

[2]*Levich Institute and Physics Department,*
City College of New York,
New York, NY 10031, USA

[3]*LMIB and School of Mathematics and Systems Science,*
Beihang University,
Beijing, 100191, China

[4] *Biomedical Engineering Department,*
City College of New York,
New York, NY 10031, USA

[5]*Departamento de Física,*
Universidade Federal do Ceará,
60451-970 Fortaleza, Ceará, Brazil

[6]*Department of Physics,*
Bar-Ilan University,
52900 Ramat-Gan, Israel

* *These authors contributed equally.*

Abstract

The probability distribution of number of ties of an individual in a social network follows a scale-free power-law. However, how this distribution arises has not been conclusively demonstrated in direct analyses of people's actions in social networks. Here, we perform a causal inference analysis and find an underlying cause for this phenomenon. Our analysis indicates that heavy-tailed degree distribution is causally determined by similarly skewed distribution of human activity. Specifically, the degree of an individual is entirely random - following a "maximum entropy attachment" model - except for its mean value which depends deterministically on the volume of the users' activity. This relation cannot be explained by interactive models, like preferential attachment, since the observed actions are not likely to be caused by interactions with other people.

INTRODUCTION

Millions of people edit Wikipedia pages, however, in average we find that only 5% contribute to 80% of their content. Such heterogeneous level of activity is reminiscent of the well-known and widely applicable law postulated by Pareto [1], which states that 80% of the effects are induced by 20% of the causes. The example of Wikipedia users reported here highlights how heterogeneous the activity of their users are, with both, activity as well as degree following a power-law distribution.Indeed, heavy-tailed distributions following a power-law have been observed in variety of social systems ever since Pareto reported his observation of the extreme inequality of wealth distribution in Italy back in 1896 [1]. In recent years, due to ubiquitous computerization, networking and obsessive data collection, reports of heavy-tailed distributions have almost become a routine [2–6]. Following simple distributions such as those of wealth, and income [7], certain structural properties of social systems were also found to be heavy-tailed distributed. More specifically, distribution of the number of ties of a person (degree) has been shown to fall in this group for vast and still growing number of social networks [8, 12]. Power-law degree distributions, called scale-free [8], represent one of the three general properties of social networks (short distances and high clustering being the other two [13]). A power-law degree distribution is not only the least intuitive and surprising property, but also is the most well-studied and debated feature of networks since extensively found in the late 90s [8, 14].

Immediately following the empirical measurements, a number of plausible models aiming at explaining the emergence of these distributions have been proposed [8–11, 15, 16]. Many models reproduce heterogeneous connectivity by amplifying small differences in connectivity – frequently stochastically emerging – using some kind of multiplicative process or "preferential attachment" [8–11, 15–18]. Other models propose different optimization strategies leading to scale-free [19–21]. A common attribute of all these models is that fat-tailed distributions emerge out of some kind of interaction between the basic system's elements. In fact, the question is not whether there exists a mechanism that could produce scale-free networks similar to the ones observed, but which of the many mechanisms suggested are more likely to actually play a significant role in each network formation.

The data presented here suggests that there is a different underlying cause for heavy-tailed degree distributions which does not involve interactions between people. We investi-

gate distinct social networks focusing on the relationship between users' activity and degree, specifically, the number of posts, messages, or actions of a user, i.e. *activity* and the number of user establishing a link with her/him, i.e. the incoming degree, or *degree*, for short. Both, degree k in the social network, and the activity A of a user, exhibit power-law distributions $P(k) \sim k^{-\gamma_k}$, and $P(A) \sim A^{-\gamma_A}$, where γ_k and γ_A are the scale-free degree and activity exponents, respectively. Positively skewed distributions of human activity were recently reported in [22, 23] and we extend this result here for a number of datasets. More importantly, in all instances we find that activity causally determine degree of the same user, suggesting that the broad distribution of one, could result from the broad distribution of the other. It is important to note that the studied actions are not likely to be driven by interaction with other people. Activity and degree, as measured here, are taken from two different networks developed by the same pool of users, and so there is no reason to expect that they should depend on each other in some trivial fashion. Surprisingly, however, the number of potential followers of a user (degree distribution) appears to be entirely random except for its mean value, which is tightly controlled by the volume of activity of that user. Our observations convincingly point at the intrinsic activity of people as the driving force behind the evolution of the examined social systems and particularly the heterogeneity in user connectivity. The observed degree distribution in social systems may merely be a manifestation of the similarly wide distribution of human activity related to the system construction. These wide distributions in social collaborative networks cannot be explained by interactive model since the observed actions are not likely to be caused by actions of other people.

RESULTS

Network construction

We have analyzed activity of individuals over time collaboratively working on construction of extensive electronic data sets: Wikipedia in four different languages (*http://www.wikipedia.org*), and a collaborative news-sharing web-site (*http://www.news2.ru*). These datasets represent various domains of human activity and contain records of a vast number of individual user contributions to the collaboratively generated content (see Method). For each person, we analyze two properties defined in two independent layers: activity and degree. For

instance, in Wikipedia, the activity performed by users includes posting of new material and discussions about them. This is the activity layer. Simultaneously, by tracing users contributing to other users' personal or talk pages, we recover the underlying network of Wikipedia contributors' personal communication or social network. The resulting network reliably represents actual interactions of Wikipedia users [24–26] and thus defines the social network layer. The number of incoming connections, i.e. others reaching out to the user in this network represents the degree. In principle, activity and degree as defined here are unrelated. Similarly, news2.ru posses the same two-layer structure of activity and degree (see Method).

Analysis of activity and degree distribution

We start by analyzing the distributions of various types of activities performed by users in these systems. Very few of the most active users perform the vast majority of work so that the activity levels frequently span five orders of magnitude (Fig. 1a,b). For instance, when analyzing the activity to a given Wikipedia page, only 5% of users contribute 80% of the edits (Fig. 5 in Method). This surprising result is similar to the $80-20$ rule postulated by Pareto [1] to describe the unequal distribution of wealth. Indeed, a power-law faithfully characterizes the activity distributions in Fig. 1. The exponent of the activity distribution for Spanish language Wikipedia is $\gamma_A = 1.752 \pm 0.005$ (Fig. 1a), while the activity distribution for voting in stories in News2.ru is $\gamma_A = 1.88 \pm 0.04$ (Fig. 1b, detailed fitting procedure in Method [32, 33])

The activity distributions in Fig. 1a represent the number of users as a function of the number of Wikipedia edits in four languages. Interestingly, different populations performing similar activity in separate instances of similarly-built social systems exhibit identical activity distributions. Figure 1b shows several different activities performed by the same population of users at the social news aggregator news2.ru. These activities differ in their complexity. We consider submission of posts to be the most difficult and time consuming of the four activity types because it typically requires the user to locate the content on-line, evaluate its quality and publish at the news2.ru web site by filling a form with multiple fields. Considering the task complexity, writing comments is arguably easier task than posting. There are on average nearly three comments per every published post. These two

content-generating tasks are followed by ranking of posts and comments. The differences in the underlying complexity of the task seem to explain the difference in the range and slope of the observed distributions plotted in Fig. 1b.

We further observe the social networks emerging in each of these systems. These networks serve different functions. In Wikipedia they arise due to the direct interaction required to coordinate common tasks. In particular, we derive social networks from the record of edits of personal user pages by other users - a common way of personal communication in Wikipedia (the web site rules forbid activity-related confidential communication between its editors). In news2.ru the social network emerges through declaration of personal attitudes - a user may indicate that he/she likes, dislikes or is neutral to any other user. Another social network arises from a set of explicit (directed) declarations of friendship between news2.ru users. Figure 1 c and d present the degree distributions in these networks. Broad distributions are measured and present in each system, suggesting a scale-free behavior in their degree distribution. The exponent of the degree distribution for Spanish Wikipedia is $\gamma_k = 1.92 \pm 0.01$ (Fig. 1c), and for the degree distribution in News2.ru is $\gamma_k = 2.11 \pm 0.08$ (Fig. 1d).

Dependence between activity and degree

The present data suggest a simple explanation of the origin of degree distributions. We first observe that the number of incoming links aggregated by a person in all these social networks is highly correlated to the individual's activity. The correlation between the degree and the activity measurements is presented in Table I. It is measured here as the correlation of the log-values to capture the gross relationship of these two variables across different orders of magnitude. More importantly, the dependence analysis below suggests that the broad distribution of activity is the driving force of scale-free degree as will be discussed next.

It is important to emphasize that in order to avoid direct and rather obvious correlation between different aspects of activity of the same person, we test the correlation of individual's activity to her degree determined by actions of his/her followers rather than his/her own. It is possible that these actions are driven by reciprocity, i.e., a person is simultaneously active in the community and in constructing her social network inspiring others to link back

to her.

To determine the precise nature of the (k, A) relationship, we analyze the joint distribution of degree and activity, $p(k, A)$ (Fig. 2a). We find that the mean degree μ_k for a given level of activity follows a smooth monotonic function of A (Fig. 2b), whereas the opposite is not true, i.e., the mean activity μ_A does not seem to be tightly determined by degree (Fig. 2c). A similarly tight relationship exists for the standard deviation of the degree distribution σ_k for specific values of the activity (Fig. 2d), but, again, the reverse is not true (Fig. 2e). The conditional mean and standard deviation of degree (conditioned on activity) show a tight relationship with approximately unit slope $\sigma_k \approx \mu_k$ (Fig. 2f). However, the σ_A, μ_A values conditioned on degree are more variable (Fig. 2g). Based on these observations we hypothesize that the conditional degree distribution $p(k|A)$ may be scale invariant with scale μ_k entirely determined by activity: $\mu_k = f(A)$. Here, this functional dependence of scale can be estimated as the mean activity for a given A: $\mu_k = f(A) \approx \text{mean}(k|A)$. Indeed, we observe that the conditional degree distribution appears to follow a geometric distribution for all μ_k:

$$p(k|\mu_k) = (\mu_k - 1)^{(k-1)} \mu_k^{-k}. \tag{1}$$

This theoretical distribution provides a remarkably accurate fit to the first two sample moments of degree for a given level of activity as shown in Fig. 3. We plot the standard deviation σ_k versus mean degree μ_k for given activity for four Wikipedia databases. The curves follow a smooth, monotonically increasing functional form which is almost identical for all datasets (as one would expect for activity conditioning degree). When the analysis is repeated for activity conditioned on degree the variables do not appear to follow a tight relationship.

The tight relationship between σ_k versus μ_k conditioned on activity follows asymptotically a straight line with unit slope, which follows exactly the geometric distribution Eq. (1). In Fig. 3, we compare the data to the analytic relationship between mean and standard deviation for geometric distribution Eq. (1): $\mu = \frac{1}{p}$ and $\sigma = \sqrt{\frac{1-p}{p^2}}$, where p is the parameter of geometric distribution. The data fit this theoretical curve surprisingly well for the four displayed languages of Wikipedia ($r^2 = 0.8889$ in average).

Dependence Hypotheses

The previous findings can be understood with the following hypothesis H1: $A \to k$, activity deterministically affects the mean degree, but degree is otherwise random (Fig. 4a). Note that for positive discrete variables – like the degree – with a given mean, the highest entropy or least informative and most random distribution is achieved by the geometric distribution as we find above [27]. The geometric distribution is analogous to exponential distribution in statistical mechanics, which maximizes entropy for continuum variables with fix mean. We also tested the inverse hypothesis H2: $k \to A$, degree deterministically affects mean activity, $\mu_A = g(A) \approx \text{mean}(A|k)$, and activity is otherwise random.

The goodness-of-fit of these two analytic models to histograms of H1: activity $\to$ degree or H2: degree $\to$ activity was measured with the χ-square statistics averaged over activity or degree respectively. The likelihood that the observed distributions match H1 or H2 was assessed using surrogate data generated with Monte-Carlo sampling to estimate the chance occurrence of these averaged χ-square values. The results for the Spanish language Wikipedia data indicate that we cannot dismiss the correctness of H1 (Fig. 4b) with a confidence of higher than 95% ($p = 0.23$) but that H2 can be soundly dismissed (the chance of the corresponding χ-square value occurring at random is $p < 10^{-5}$). The same is true for all other datasets (see Table I). In all datasets the likelihood of H1 is several orders of magnitudes larger than H2 and thus we accept model H1, which states that activity determines degree.

Given the explicit model of a geometric distribution for $P(k|A)$ of hypothesis H1, and the observed power-law distribution for activity, $P(A) \sim A^{-\gamma_A}$, one can explicitly derive the expected degree distribution.The conditional degree distribution closely matches a geometric distribution (Fig. 3). For large mean values, say $\mu_k > 10$, it can be very well approximated by its continuous equivalent, the exponential distribution i.e. $P(k|A) = \frac{1}{\mu_k} e^{-\frac{k}{\mu_k}}$. Therefore:

$$P(k) = \int dA P(k|A)P(A) \sim \int dA \frac{1}{\mu_k} e^{-\frac{k}{\mu_k}} A^{-\gamma_A} \tag{2}$$

$$= \int dA A^{-\delta} e^{-\frac{k}{A^\delta}} A^{-\gamma_A}, (u = \frac{k}{A^\delta}), \tag{3}$$

$$= -\int \frac{du}{\delta} u^{\frac{\gamma_A - 1}{\delta}} e^{-u} k^{\frac{1-\gamma_A}{\delta} - 1} \tag{4}$$

$$\sim k^{\frac{1-\gamma_A}{\delta} - 1} \sim k^{-\gamma_k}. \tag{5}$$

Thus the exponent is predicted to be

$$\gamma_k = 1 + \frac{\gamma_A - 1}{\delta}. \tag{6}$$

where δ defines $\mu_k \sim A^\delta$ for large A as shown in Figure 2b. The observed exponents γ_k closely follow these predicted exponents for all datasets (Table I).

DISCUSSION

The causal inference argument provided here is borrowed from ideas recently developed in causal inference [28–30]. There, a deterministic functional dependence of cause on mean effect is hypothesized and deviations from this mean effect are assumed to have fixed standard deviation but to be otherwise random. With two variables for which one wishes to establish causal direction, the model is evaluated in both directions and the more likely one is postulated to indicate the correct causal dependence, as we have done here. This approach has been demonstrated to give the correct causal dependence for a large number of known causal relationships [31], and theoretical results indicate that there is only an exceedingly small class of functional relationships and distributions for which this procedure would give the incorrect answer. Such an identifiability proof does not yet exist for the present case where the standard deviation is not constant. Nevertheless, our explicit model of a deterministic effect of human activity on the success of establishing social links is the simplest possible explanation for the data available to us. For a different dataset a different probabilistic model may be better suited.

The individual activity of people deterministically affects the mean success at establishing links in a social network, and the specific degree of a given user is otherwise random following a maximum entropy attachment (MEA) model. The MEA model is exemplified in Fig. 4a and consists of the following steps: Introduce a node i with q links, where q is drawn from a probability given by the activity of the node. The activity has an intrinsic power-law distribution. Then, link the q links at random following maximum entropy principle with the concomitant geometric distribution $P(k|\mu_k)$. This mechanism contrasts with the preferential attachment mechanism [8, 15–18] where each link attaches to a node with a probability proportional to the number of links of that node. A possible mechanism by which a geometric distribution could arise is based on the notion of "success". In this model,

the activity of users aims to achieve a specific outcome (a Wikipedia project), and each new incoming link can aid in achieving this desired outcome; once the goal is achieved the user stops collecting links. The probability of the desired event in this model is $q = 1/\mu_k \sim A^{-\delta}$. Hence, those users working so very hard may have an exceedingly unlikely event they are aiming for. But eventually, they too will succeed, and will turn their attention away from the on-line social network.

The present data indicates that degree distribution is maximally random except for what can be determined solely from the volume of a user's activity. Does this mean that the precise content of a user's actions (the meaning and quality of the edits in Wikipedia, messages, etc) is immaterial in determining his/her success in establishing relationships? One can only hope that small deviations from this maximum entropy attachment model will become more pronounced with increasing data-set sizes, which can then point us to the benefits of well thought out and carefully executed actions, specially in specialized large-scale collaborative projects like Wikipedia.

Whether the dynamics of preferential attachment is consistent with the maximum entropy distribution of degree remains to be established. What is certain is that distributions of levels of activities in all tested populations are heavily heavy-tailed indicating highly varying level of involvement of users in collaborative efforts. We showed here that this fact alone is sufficient to produce the heavy-tailed distribution of degree observed throughout social networks. Therefore, previous interactive models may not be necessary. The present result shifts the burden of proof to explaining the origin to the incredible diversity in human effort observed here spanning five orders of magnitude.

METHOD

Datasets

The number of actions contained in the datasets range from hundreds of thousands to hundreds of millions of user actions. From the editing on Wikipedia, to the votes, to commentaries on News2.ru, these actions represents different and natural underlying dynamics of social networks, since they range from collaborative interaction (Wikipedia) to discussions about different interesting of human behavior (New2.ru), which are intrinsic properties of

the social nature of the web.

We have collected details about user activity in the Wikipedia project and reconstructed the underlying social network. In addition to the widely used term and category pages, Wikipedia provides special pages associated with specific contributing authors and discussion (talk) pages maintained alongside each of these pages. These user pages are widely used by Wikipedia contributors for coordination behind the scenes of the project. In fact, interaction via user and discussion pages dominates all other communication methods. However, communication via personal user pages (and the corresponding discussion pages) differs from the topic-associated talk pages in that it is explicit person-to-person communication rather than general topic specific, usually impersonal communication. By tracing users contributing to other user's personal or talk pages, we recover the underlying network of Wikipedia contributor's personal communication. Not surprisingly, as presented in the next section, the obtained social networks show a scale-free degree distribution, typically observed in a variety of social networks analyzed so far.

The other data set is a de-identified record of activities of social news aggregator news2.ru. The record contains all actions performed by the community members over more than three years of collaborative selection and discussion of news-related content. These, user-related actions include such events as submission of news article, comments as well as preference-revealing actions such as voting for articles ("dig" or "bury", using digg.com language) and other users' comments. In addition to the trace of user activity, the data contains explicit social network layer. Each user may publicly declare his/her (positive, neutral or negative) attitude to any other user. Considering the personal flavor of the rather emotional way people interact through commentary threads, this list of attitudes when aggregated can be perceived as social network. In addition, users maintain list of friends, usually including users most favorable on them. These networks are directional, which allows to focus on the incoming links, since they can not be controlled by the target individual, but by his/her friends.

Each of these systems represents different approaches to collaborative content creation. The Wikipedia editors interact to create the same content collaboratively so that the content contributed by one user can be complemented, altered or completely removed by others. The news2.ru represents a mixed case in which the content is contributed individually, but collaboratively ranked. Given these fundamental differences in user activity and network

dynamics, the similarities between these systems reported below are particularly revealing.

Method of Power-law Fitting

To get the exponents γ_k and γ_A of power-law distribution, we present a rigorous statistical test based on maximum likelihood methods [32]. Take the degree distribution as an example. We fit degree distribution assuming a power law within a given interval. For this, we use a generalized power-law form

$$P(k; k_{min}, k_{max}) = \frac{k^{-\gamma}}{\zeta(\gamma, k_{min}) - \zeta(\gamma, k_{max})}, \tag{7}$$

where k_{min} and k_{max} are the boundaries of the fitting interval, and the Hurwitz ζ function is given by $\zeta(\gamma, \alpha) = \sum_i (\gamma + \alpha)^{-\gamma}$.

We use the maximum likelihood method, following the rigorous analysis of Clauset et al. [32]. The fit was done in an interval where the lower boundary was k_{min}. For each k_{min} value we fix the upper boundary to $k_{max} = K$, where K is the maximal degree. We calculate the slopes in successive intervals by continuously increasing k_{min} and varying the value of w. In this way, we sample a large number of possible intervals. For each one of them, we calculate the maximum likelihood estimator through the numerical solution of

$$\gamma = argmax(-\gamma \sum_{i=1}^{N} \ln k_i - N \ln[\zeta(\gamma, k_{min}) - \zeta(\gamma, k_{max})]), \tag{8}$$

where k_i are all the degrees that fall within the fitting interval, and N is the total number of nodes with degrees in this interval. The optimum interval was determined through the Kolmogorov-Smirnov (KS) test.

For the goodness-of-fit test, we use the Monte Carlo method described in [32]. For each possible fitting interval, we calculate the Kolmogorov-Smirnov statistics D for the obtained cumulative distribution function. Then we choose the interval with the minimal D as the best fitting interval and take the γ in this interval as the final result. As to the standard error estimation, we adopt the method in [32]. The standard error on γ, which is derived from the width of the likelihood maximum, is $e = (\gamma - 1)/\sqrt{n} + O(1/n)$, where n is the number of data.

Although the fitting method mentioned above is rigorous, it is suitable for fitting probability density distributions. When we fit the data $\mu_k = A^{\gamma_A}$, we use another fitting method [33].

The procedure for determining fitting interval is similar. In each fitting intervals, the fittings were done using ordinary least squares methods. The goodness of fitting was estimated through the coefficient of determination, r^2, where $0 \leq r^2 \leq 1$. The value of r^2 is used as a measure of how reliably the fitted line describes the observed points, and is often described as the ratio of variation that can be explained by the fitted curve over the total variation. We assume that any value above $r^2 \geq 0.85$ represents an accepted fitting. The final result is the average of the accepted exponent.

Users contributing to 80% of a Wikipedia page

In Fig.5, each dot represents a distinct Wikipedia project page. Horizontal axis measures the total number of edits for each project. Vertical axis represents the fraction of contributors to that project who performed 80% of edits on that project. This fraction drops fast (with power law) as the number of edits grows. This suggests that the largest projects are dominated by a few very dedicated users. Perhaps more representative are the mean values; the vertical line indicates the average edits and the horizontal line marks the fraction of users contributing 80% if the work in the average across projects (approximately 5%)

Monte-Carlo sampling for hypothesis tests

The accuracy of fit of the data to the theoretical geometric distribution is measured as the χ^2 goodness-of-fit to the conditional histogram. As an example, consider H1 for the Spanish Wikipedia data: For the theoretical distribution we use for each activity the mean degree μ_k as shown in Fig. 2b. The χ^2 value is then averaged over all activity bins shown in that figure. To test if this observed average χ^2 is consistent with chance assuming H1 we generate surrogate data following H1: For each given activity, we generate the same amount of random numbers from a geometric distribution with the same mean values, calculate the χ^2 values and again, average across activities. We draw 10^5 such samples and obtain a distribution of average χ^2 (Fig. 4b). The chance that the χ^2 for the Spanish Wikipedia data occurred by chance (p-value) is the fraction of times the surrogate data provided a value larger than the one observed (red line in Fig.4b). The analysis for H2 is analogous using the

data as shown in Fig. 2c. The resulting p-values for all datasets can be found in Table I.

[1] V. Pareto, *Cours D'economie Politique* (F. Rouge, Luzanne, 1896).

[2] B. A. Huberman and L. A. Adamic, Internet: Growth dynamics of the World-Wide Web, *Nature* **401**, 131 (1999).

[3] R. L. Axtell, Zipf distribution of U.S. firm sizes, *Science* **293**, 1818 (2001).

[4] C. Castellano, S. Fortunato, and V. Loreto, Statistical physics of social dynamics, *Rev. Mod. Phys.* **81**, 591 (2009).

[5] A.-L. Barabási, The origin of bursts and heavy tails in human dynamics, *Nature* **435**, 227 (2005).

[6] D. Rybski, S. V. Buldyrev, S. Havlin, F. Liljeros, and H. A. Makse, Scaling laws of human interaction activity, *Proc. Natl. Acad. Sci. USA* **106**, 12640 (2009).

[7] V. M. Yakovenko and J. B. Rosser, Statistical mechanics of money, wealth, and income, *Rev. Mod. Phys.* **81**, 1703 (2009).

[8] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, *Science* **286**, 509 (1999).

[9] G. Caldarelli, A. Capocci, P. De Los Rios, and M.A. Muñoz, Scale-Free Networks from Varying Vertex Intrinsic Fitness, *Phys. Rev. Lett.* **89**, 258702 (2002).

[10] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, Modeling of Protein Interaction Networks, *ComPlexUs* **1**, 38 (2003).

[11] G. Bianconi and A.-L. Barabási, Competition and multiscaling in evolving networks, *Europhys. Lett.* **54**, 436 (2001).

[12] G. Caldarelli, *Scale-Free Networks: Complex Webs in Nature and Technology* (Oxford Univ Press, Oxford, 2007).

[13] D. Watts and S. Strogatz, Collective dynamics of 'small-world' networks, *Nature* **393**, 440 (1998).

[14] M. Faloutsos, P. Faloutsos, and C. Faloutsos, On power-law relationships of the Internet topology, *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication* (ACM, New York, 1999), pp 251–262.

[15] M. E. J. Newman, The Structure and Function of Complex Networks, *SIAM Rev.* **45**, 167 (2003).

[16] M. Mitzenmacher, A brief history of generative models for power-law and lognormal distributions, *Internet Mathematics* **1**, 226 (2004).

[17] G. U. Yule, A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F.R.S., *Philos. Trans. R. Soc. Lond. B* **213**, 21 (1925).

[18] H. A. Simon, On a class of skew distribution functions, *Biometrika* **42**, 425 (1955).

[19] B. Mandelbrot, *Communication Theory, ed. W. Jackson* (Butterworth, London, 1953), pp. 486–502.

[20] R. M. D'Souza, C. Borgs, J. T. Chayes, N. Berger, and R. D. Kleinberg, Emergence of tempered preferential attachment from optimization, *Proc. Natl. Acad. Sci. USA* **104**, 6112 (2007).

[21] F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguñá, and D. Krioukov, Popularity versus similarity in growing networks, *Nature* **489**, 537 (2012).

[22] J. Leskovec and E. Horvitz, Planetary-scale views on a large instant-messaging network, *Proceedings of the 17th international conference on World Wide Web*, pp. 915–924 (2008).

[23] Perra, N., Gonçalves, B., Pastor-Satorras, R. & Vespignani, A. Activity driven modeling of time varying networks. *Scientific Reports* **2**, 469 (2012).

[24] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, Structure and evolution of blogspace, *Communications of the ACM - The Blogosphere* **47**, 35 (2004).

[25] W. H. Hsu, J. Lancaster, M. S. R. Paradesi, and T. Weninger, Structural link analysis from user profiles and friends networks: a feature construction approach, *Proceedings of the International Conference on Weblogs and Social Media*, pp 75–80 (2007).

[26] D. Liben-Nowell and J. Kleinberg, Tracing information flow on a global scale using Internet chain-letter data, *Proc. Natl. Acad. Sci. USA* **105**, 4633 (2008).

[27] F. Topsøe, Information theoretical optimization technique, *Kybernetika* **15**, 8 (1979).

[28] J. Pearl, Causal inference in statistics: An overview, *Statistics Surveys* **3**, 96 (2009).

[29] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf, Nonlinear causal discovery with additive noise models, *Proceedings of the Conference Neural Information Processing Systems*, (2009).

[30] K. Zhang and A. Hyvärinen, On the identifiability of the post-nonlinear causal model, *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence* (AUAI Press, Arlington,

2009), pp. 647–655.

[31] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf, Information-geometric approach to inferring causal directions, *Artificial Intelligence* **182**, 1 (2012).

[32] A. Clauset, C. R. Shaliz, and M. E. J. Newman, Power-law distribution in empirical data, *SIAM Rev.* **51**, 661 (2009).

[33] L. K. Gallos, P. Barttfeld, S. Havlin, M. Sigman, and H. A. Makse, Collective behavior in the spatial spreading of obesity, *Sci. Rep.* **2**, 454 (2012).

ACKNOWLEDGMENTS

We thank G. Khazankin, Research Institute of Physiology SB RAMS for kindly providing access to invaluable data on news2.ru user activity. The research is supported by NSF Emerging Frontiers, ARL, FP7 project SOCIONICAL and MULTIPLEX, CNPq, CAPES, and FUNCAP.

AUTHOR CONTRIBUTIONS

H.A.M., S.H. and J.S.A. designed research. L.M. prepared data. L.M., S.P., L.C.P. and S.D.S.R. analyzed the data. All authors wrote, reviewed and approved the manuscript.

ADDITIONAL INFORMATION

Competing financial interests: The authors declare no competing financial interests.

Networks	$r_{\log}$	p_{H1}	p_{H2}	δ	γ_A	γ_k	predicted γ_k
Spanish	0.64	0.23	$< 10^{-5}$	0.79 ± 0.02	1.752 ± 0.005	1.92 ± 0.01	1.95 ± 0.03
Italian	0.69	0.11	$< 10^{-5}$	0.70 ± 0.04	1.620 ± 0.004	1.85 ± 0.01	1.88 ± 0.05
Russian	0.69	0.13	$< 10^{-5}$	0.68 ± 0.03	1.618 ± 0.007	1.89 ± 0.01	1.91 ± 0.05
Hebrew	0.77	0.16	$< 10^{-5}$	0.67 ± 0.04	1.574 ± 0.008	1.80 ± 0.01	1.85 ± 0.05
Story	0.64	0.10	$< 10^{-5}$	0.79 ± 0.08	1.98 ± 0.04	2.11 ± 0.08	2.2 ± 0.1
Comment	0.68	0.37	$< 10^{-5}$	0.72 ± 0.09	1.88 ± 0.05	2.11 ± 0.08	2.2 ± 0.2
Story Vote	0.65	0.05	$< 10^{-5}$	0.70 ± 0.08	1.88 ± 0.04	2.10 ± 0.08	2.3 ± 0.2
Comment Vote	0.59	0.26	$< 10^{-5}$	0.71 ± 0.09	1.85 ± 0.09	2.1 ± 0.2	2.2 ± 0.2

TABLE I. Statistics for different datasets. The log-correlation $r_{\log}$ between the user's activity and his/her degrees in Wikipedia and News2.ru is displayed in the first column. p_{H1} and p_{H2} are p-values for hypotheses H1 and H2, respectively. δ is the exponent for $\mu_k \sim A^{\delta}$, while γ_A and γ_k are the power law exponents of activity and degree distribution obtained by fitting the data. The predicted γ_k results from scaling relation as detailed in the text.

Fig. 1. Probability distribution of activities and degree. (a) Probability density function of Wikipedia contributors as a function of the number of performed page edits in four languages. (b) Probability density function of news2.ru for five different activities.Lines indicate power-law fitting for Spanish and Stories with the maximum likelihood methods. (c) Probability distribution of degree for social networks as a function of number of links between Wikipedia contributors. Degree represents the number of links other users establish with a given user. (d) Distribution for networks of relationship (positive/negative) between users of news2.ru web portal and users' friendships.

Fig. 2. Analysis of joint distribution of activity and degree. (a) Scatter plot of degree and activity for each user in Wikipedia Spanish dataset. (b) Mean degree μ_k for given activity. (c) Mean activity μ_A for given degrees. (d) Standard deviation of degree σ_k for given activity. (e) σ_A for given degree. (f) Relationship between standard deviation of degree σ_k and the mean value μ_k for given activity. Inset is the theoretical fit of geometric distributions for Spanish Wikipedia. (g) σ_A versus μ_A for given degree.

Fig. 3. Test of "maximum entropy attachment model" via the geometric distribution. Theoretical relationship of mean and standard deviation for geometric distribution (solid curve) and data points for Wikipedia in four languages.

Fig. 4. Causal hypotheses and test result. (a) Schematic diagrams for hypotheses H1 and H2. H1: Mean degree is determined by activity through function $\mu_k = f(A)$. Then degree is random distributed according to the conditional probability distribution $P(k|\mu_k)$. H2 is the other way around. (b) and (c) Results of Monte-Carlo simulation with 10^5 samples following H1 and H2 for the Spanish Wikipedia data. The vertical red lines show the goodness-of-fit χ^2 of the actual data to H1 and H2, respectively. The empirical analysis clearly favors H1 over H2.

Fig. 5 (color online). Users contributing to 80% of a Wikipedia page.

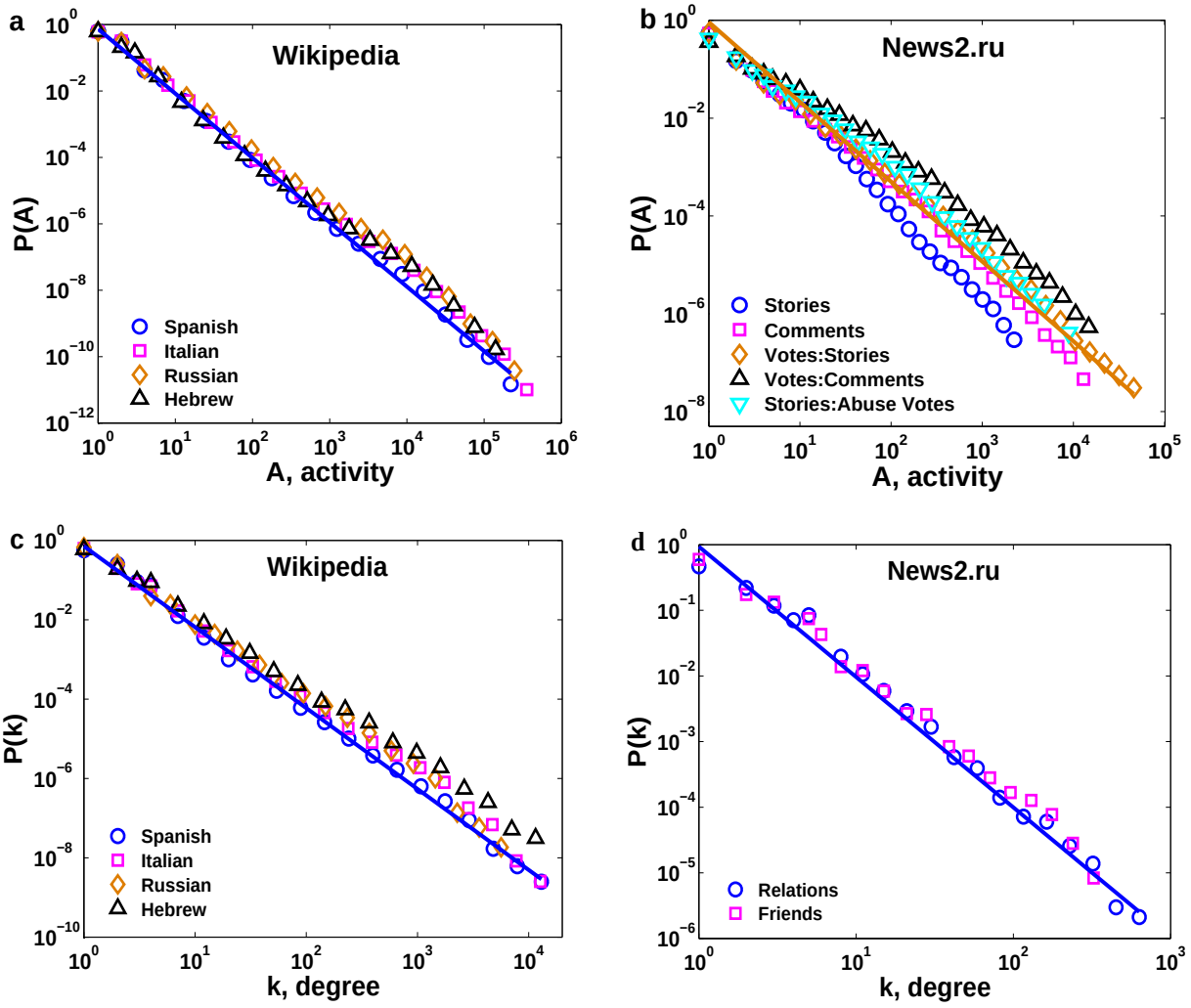
a
Wikipedia
P(A)
A, activity
Spanish
Italian
Russian
Hebrew
b
News2.ru
P(A)
A, activity
Stories
Comments
Votes:Stories
Votes:Comments
Stories:Abuse Votes
c
Wikipedia
P(k)
k, degree
Spanish
Italian
Russian
Hebrew
d
News2.ru
P(k)
k, degree
Relations
Friends

FIG. 1.

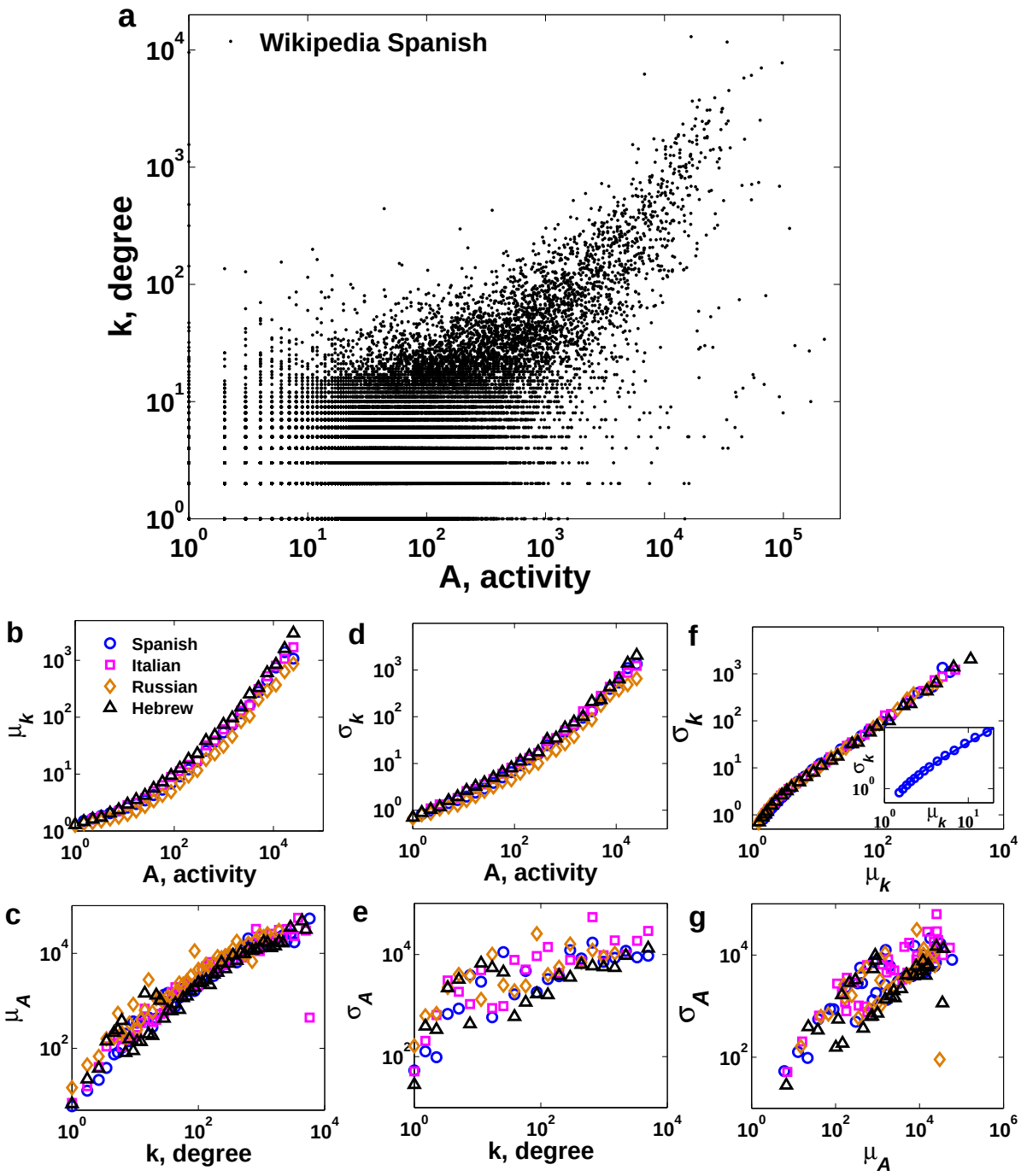
a
Wikipedia Spanish
k, degree
A, activity
b
Spanish
Italian
Russian
Hebrew
A, activity
d
A, activity
f
c
k, degree
e
k, degree
g

FIG. 2.

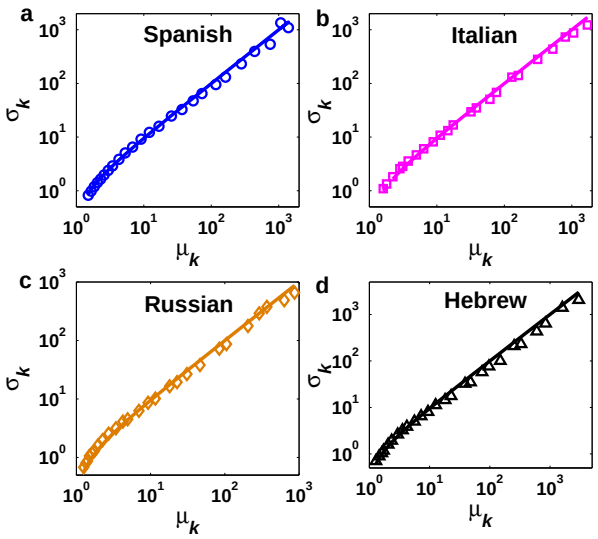
a
Spanish
b
Italian
c
Russian
d
Hebrew
σ_k
μ_k
10^0
10^1
10^2
10^3

FIG. 3.

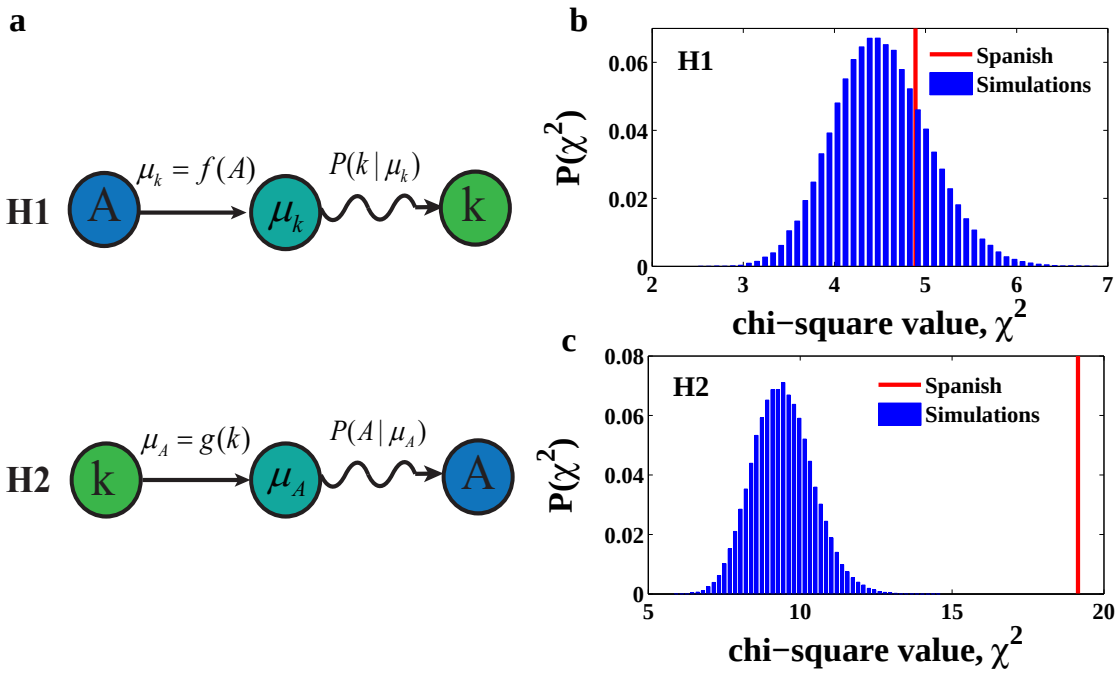
a
H1
A
$\mu_k = f(A)$
μ_k
$P(k\,|\,\mu_k)$
k
H2
k
$\mu_A = g(k)$
μ_A
$P(A\,|\,\mu_A)$
A
b
H1
Spanish
Simulations
$P(\chi^2)$
0.06
0.04
0.02
0
2
3
4
5
6
7
chi−square value, χ^2
c
H2
Spanish
Simulations
$P(\chi^2)$
0.08
0.06
0.04
0.02
0
5
10
15
20
chi−square value, χ^2

FIG. 4.

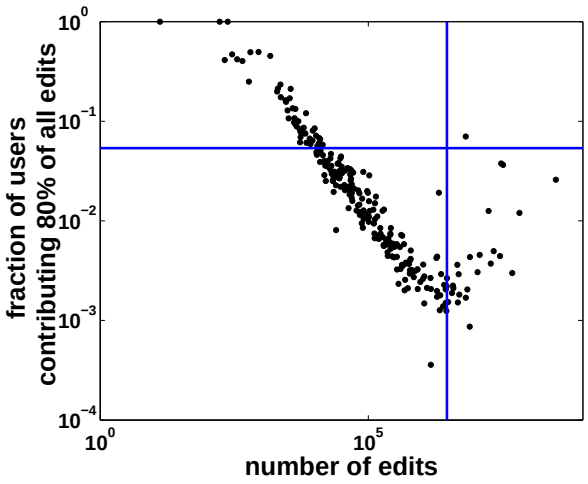
fraction of users
contributing 80% of all edits
number of edits
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^0
10^5

FIG. 5.